https://doi.org/10.29042/2020-10-5-38-43

Helix (2020) 10 (5): 38-43

A Wavelet Based Concatenation Algorithm for Gujarati Speech Synthesis

✉ [1] Priyanka Vishwas Gujarathi, [2] Sandip Raosaheb Patil
[1] JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune-411033
jspmpriyanka@gmail.com
[2] Bharati Vidyapeeth College of Engineering for Women, Dhankawadi, Pune-411043
srpatil44@gmail.com

Received: 24th August 2020, Accepted: 05th September 2020, Published: 31st October 2020

Abstract
Speech is used to express information, emotions, and feelings. The goal of Text to speech system (TTS) is the automatic conversion of written text into corresponding speech signal. TTS system is widely useful for the people with disabilities and very useful tool for visually challenged person. TTS synthesizers are available in many Indian languages along with English. It has been observed that people feel more comfortable in hearing their own native language than other language. In this research work Guajarati language is used and small footprint of database is created. Gujarati (ગુજરાતી) is an Indo-Aryan language spoken by the people of Gujarat. Indian languages are syllable centred, where pronunciations are mainly based on syllables. A Syllable can be the best unit for Indian language Speech synthesis systems. In this paper Wavelet-Domain algorithm is used for concatenation of extracted syllables. Different syllables are concatenated to get natural sounding words. Subjective listening test Mean Opinion Score (MOS) is carried out with list of questionnaires.

Keywords
Text to Speech (TTS), Syllable, Wavelet Based Concatenation, Gujarati, Concatenation CTTS.

Introduction

The speech synthesis systems are developed for the many Indian languages such as Hindi, Tamil, Telugu, Kannada, Bengali, Marathi etc. Speech corpus building task is a different than that of the English speech corpus task in Indian languages. For building natural sounding speech synthesis system, it is important that the text processing component must produce an appropriate sequence of phonemic sound units corresponding to an arbitrary (raw) input text. Phones, diphones, triphones, syllables and many more are the basic unit for speech synthesis systems but Indian languages are syllable centered, where pronunciations are typically based on syllables [1]. It is not practically feasible to cover all possible syllables in language with all possible texts. Hence some discontinuity is observed at the time of concatenation. Hence prosody modification of syllable is required before concatenation to match fundamental frequency and intensity of signal. In order to handle all these challenges, rigorous training of database is needed and can focus on prosody and speech segmentation (Labeling.) For Prosodic manipulation TD-PSOLA method is used to maintain naturalness of a prerecorded voice in CTTS. Smoothing of spectral discontinuities in the FDPSOLA representation leads to poor results [9].Parametric approaches such as "source filter"-LPC-PSOLA tends to lose detail information and so decrease in naturalness. So Wavelet-Domain algorithm is used to avoid both shortcomings [9]. The main requirement in speech synthesis is high quality and natural sounding speech. For speech synthesis different approaches are used such as articulatory text-to-speech (TTS) synthesis, Concatenative synthesis, formant based, statistical synthesis (STTS). Some Audible discontinuities (sound glitches) are resulted in case Concatenative speech synthesis system where as in case of Statistical TTS (STTS) systems, no discontinuities in synthesized speech output. But STTS produces very lower quality it gives muffled sounding effect than CTTS. So hybrid TTS system is proposed to get advantages of both synthesis methods[2].

Methodology

Concatenative Speech Synthesis:

In this, prerecorded speech signals are segmented to extract syllables and concatenated to synthesize the desired speech signal according to input. Large quantities of speech waveforms supposed to be stored in the speech corpus database to cover maximum syllables. In this work Speech Database of 1000+ Guajarati words are created and processed further. Gujarati speech database is created and collected from a female artist used as speech corpus. Every speech sound unit has many instances with changeable context and prosodic situations. As the natural waveforms are concatenated, the quality of the synthesized speech will be extremely close to natural speech.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Steps involved:

1. Conversion of Guajarati word to English word using unicode conversion table for Guajarati language and Mapping done between Gujarati unicode to English word. Different Segmentation methods are already implemented to get syllables. Segmentation must be done accurately to get natural output. After segmentation extracted syllables are stored. Gujarati word is converted into English word using Unicode conversion table. Sample Guajarati input text **આધાશીશી** and corresponding English character text **AADHASHISHI** is shown in table:

	Unicode		**English Char**	**Gujarati Char**
Word આધાશીશી to AADHASHISHI	A86	2694	AA	આ
	AA7	2727	DH	ધ
	ABE	2750	A	◌ા
	AB6	2742	SH	શ
	AC0	2752	I	◌ી
	AB6	2742	SH	શ
	AC0	2752	I	◌ી

Table 1: Word આધાશીશી to AADHASHISHI

2. After Unicode conversion syllables are extracted using MFDFA[7] syllable segmentation technique and Gaussian based segmentation techniques implemented in matlab. Syllables are stored for further processing. 1000+ words are processed, Sample Gujarati words with syllables are given.

Sr No	**Gujarati Word**	**English Word**	**Syllable - l**	**Syllable- 2**	**Syllable - 3**	**Syllable - 4**	**No of syllables**
1	આદર્શ	aadarrsh	aa	dar	sha		3
2	આધાશીશી	aadashishi	aa	dha	shi	shi	4
3	આદિ	aadi	aa	di			2
4	આગાહી	aagahi	aa	ga	hi		3
5	આગળ	aagal	aa	gal			2

Table 2: Different Guajarati Word to English Word with Syllables

3. Steps involved in wavelet concatenation algorithm: Discrete Wavelet Transform (DWT) Coefficients are calculated for given input and suitable coupling points are calculated for smoothing irregularities in the spectral shape then Inverse Discrete Wavelet Transform is carried out.

4. Finding coupling point between two syllables: For calculation of the best possible concatenation points: Considering last frame of first syllable and first frame of second syllable. We use some of distances of spectral envelopes of adjacent frames.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

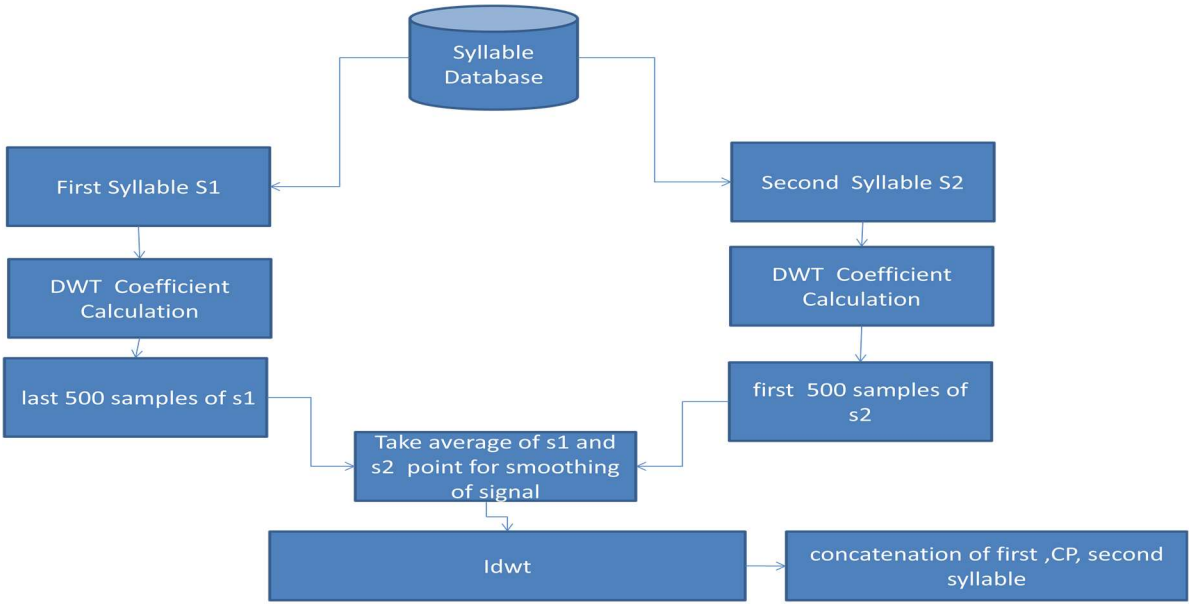

Figure 1: Syllable Concatenation Using Wavelet Concatenation Technique

Results and Discussion

1. In this concatenation of word "aagam" is shown. So from database two syllables are extracted according to requirement then DWT coefficients and coupling points are calculated for concatenation. Figure shows two syllables "aa" and "gam" it will give word "aagam" (Word=Syllable1+Syllable2)

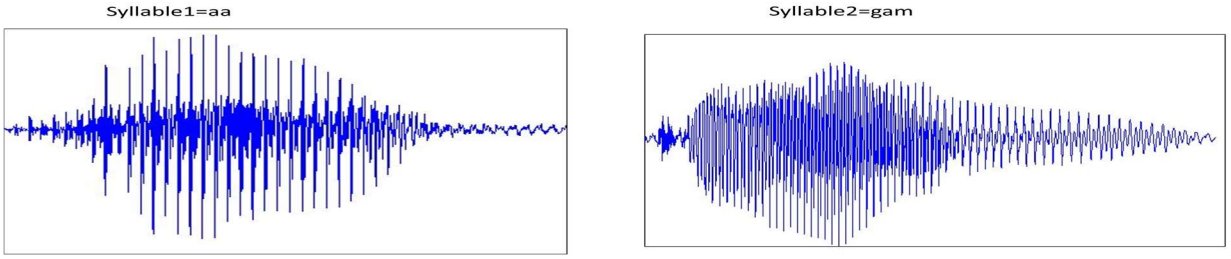

Figure 2: Syllable 1-aa Syllable 2=gam

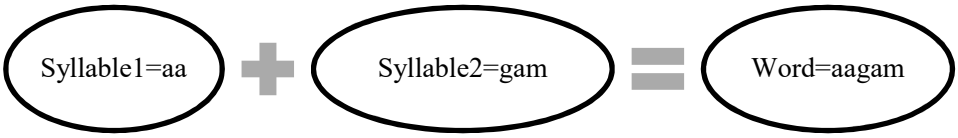

Figure 3: Word Formation from Syllables

Calculation of DWT coefficient for syllable "aa" is shown in figure. Approximation coefficients which contains low frequency components and Detail coefficients which contains high frequency components are shown in figure similarly we can calculate DWT coefficient for syllable "gam".

In DWT, A=Approximation coefficients which contains low frequency components and D=Detail coefficients which contains high frequency components.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

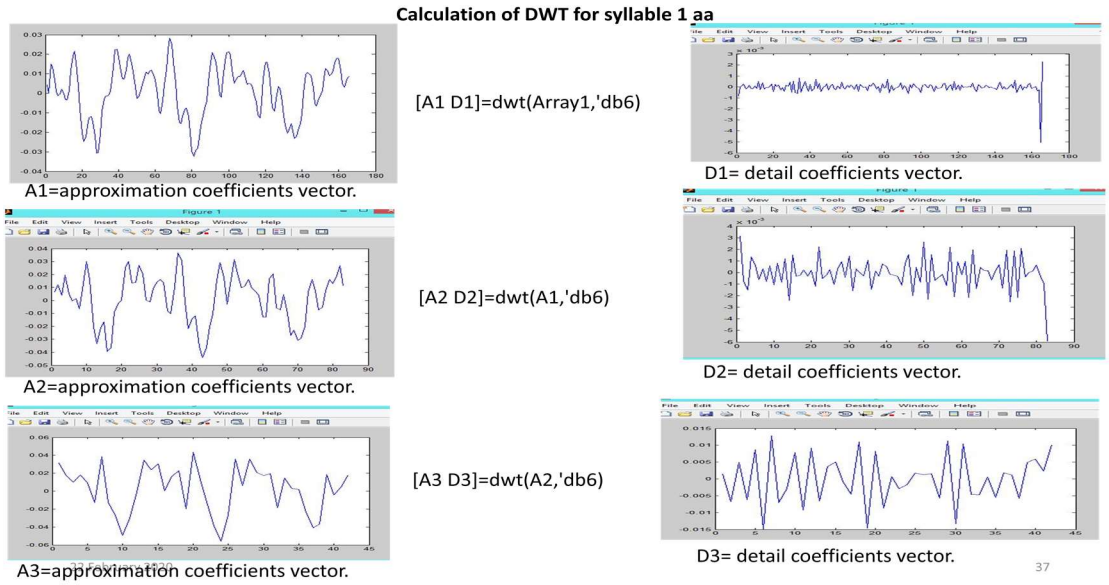

Figure 4: Calculation of DWT Coeff for Syllable1 =aa

2. Wavelet based concatenation for word: "aagam": આગમ

Wavelet coefficients are calculated for syllable aa and gam.Approximation and detail coefficients are calculated. In fig 1 shows that last 500 samples of syllable1 and first 500 samples of syllable2 are considered, then average is calculated(CP) then inverse DWT is calculated and concatenation done for syllable1 ,calculated concatenation point and syllable2 to get word aagam.

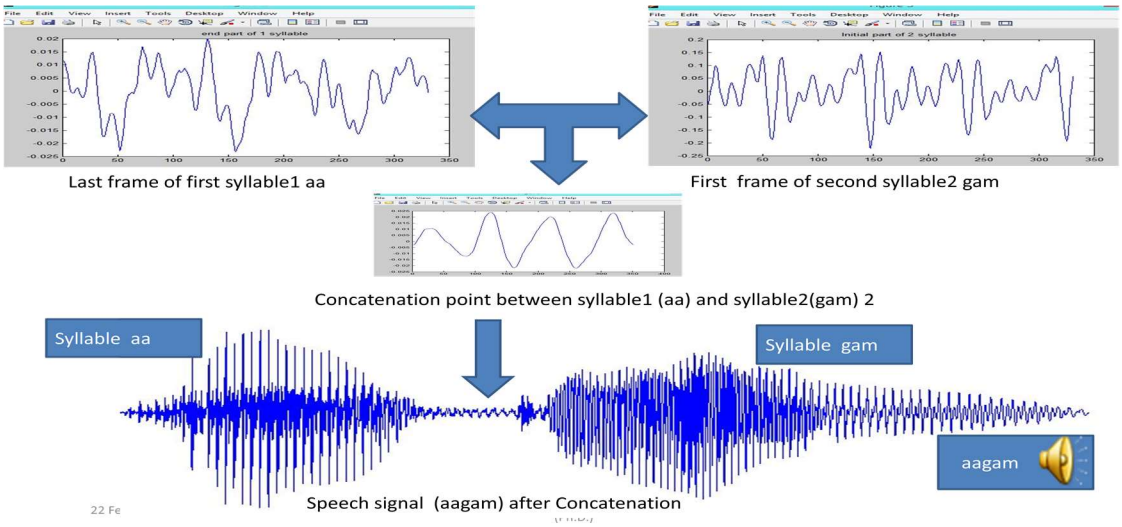

Figure 5: Concatenation of Two Syllables

3 For Validation Subjective listening test is carried out with list of questionnaires. There were 2 types of questions shown here, using a 5 points scale for 20 words and 10 listeners.

3.1 Voice pleasantness: In this test voice signal is described in terms of voice pleasantness. Subjective listening test carried out with 10 listeners and nearly 4.5 score is obtained out of 5 which means voice was pleasant while listening.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

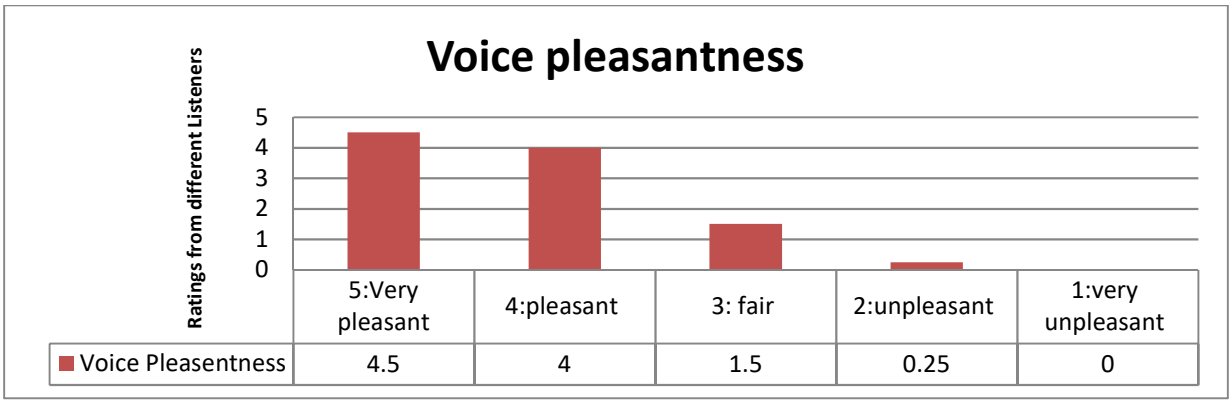

Figure 6: Voice Pleasantness

3.2 Articulation: This parameter is used to check is voice is distinguishable properly or not. Subjective listening test is carried out with 10 listeners and nearly 4.75 score is obtained out of 5.It indicates that voice signal was clear enough.

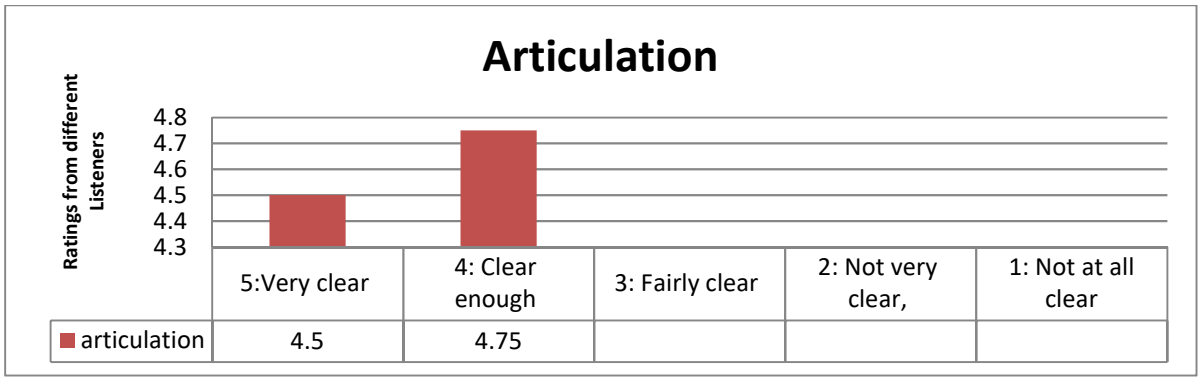

Figure 7: Articulation

Conclusion and Future work

In this paper wavelet concatenation algorithm is implemented for Guajarati speech database.Basically concatenation means stringing together different speech segments from prerecorded speech database. Generally Concatenative synthesis can generate natural sounding speech signal. In this paper, Conversion of Guajarati word to English word (unicode conversion) and mapping is done. Coupling points are calculated to concatenate two syllables or more syllables according to input. Finally concatenated output is obtained. Subjective listening test is carried to find Articulation, Speaking Rate, voice pleasantness. For future work there is requirement of rigorous training of database with different speakers to cover all possible syllables with different accents.

References

1. Kiruthiga S & Krishnamoorthy K, “Design Issues in Developing Speech Corpus for Indian Languages – A survey”, 2012 International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, Coimbatore, INDIA, 2012.
2. StasTiomkin, David SlavaShechtman, and ZviKonsMalah, “A Hybrid Text-to-Speech System That Combines Concatenative and Statistical Synthesis Units” IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 5, JULY 2011
3. Swati .Talesara Hemant A. Patil ; Tanvina Patel ; Hardik Sailor “A Novel Gaussian Filter-Based Automatic Labeling of Speech Data for TTS System in Gujarati Language” IEEE 2013 International Conference on Asian Language Processing
4. S. P. Kishore and A. W Black , "Unit Size in Unit Selection Speech Synthesis," in EUROSPEECH, Geneva
5. Romain Prudon & Christophe d’Alessandr, “A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation”, BP133 - Universit´e Paris XI, F91403 Orsay, France, ISCA ITRW ,speech syntehsis, Scotland 2001

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

6. Prof Dr.S.D.Apte,"Speech and Audio Processing" WILEY INDIA EDITION
7. Espen A.F. Ihlen, "Introduction to multifractal detrended fluctuation analysis in Matlab", Department of Neuroscience, Norwegian University of Science and Technology, Trondheim, Norway
8. N.P. Narendra , K. Sreenivasa Rao , Krishnendu Ghosh , Ramu Reddy Vempada, Sudhamay Maity, "Development of syllable-based text to speech synthesis system in Bengali".
9. M. Holzapfel, R. Hoffmann, H.Höge," A Wavelet-Domain PSOLA Approach", COCOSDA workshop,1998
10. Soumya Priyadarsini Panda , Ajit Kumar Nayak , "A waveform concatenation technique for text-to-speech synthesis", Int J Speech Technol (2017) ,Springer

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495